

DualFaceNet: augmentation consistency for optimal facial landmark detection and face mask classification

Kritaphat Songsri-in¹, Munlika Rattaphun¹, Sopee Kaewchada², Somporn Ruang-on³

¹Department of Computer Science, Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Tha Ngio, Thailand

²Department of Information Technology and Digital Innovation, Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Tha Ngio, Thailand

³Department of Creative Innovation in Science and Technology, Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Tha Ngio, Thailand

Article Info

Article history:

Received Oct 11, 2023

Revised Feb 16, 2024

Accepted Feb 28, 2024

Keywords:

Consistency loss

Deep learning

Face landmark detection

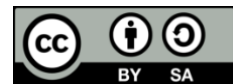
Face mask classification

Multi-task learning

ABSTRACT

In an era where face masks are commonplace, facial recognition faces new challenges and opportunities. This study introduces DualFaceNet (DFN), a cutting-edge neural network that efficiently combines facial landmark detection with mask classification. Benefiting from multi-task learning (MTL) and enhanced with a unique consistency loss, DFN outperforms traditional single-task models. Tests using the reputable 300W dataset and a face mask dataset showcase DFN's strengths: a significant reduction in landmark error to 5.42 and an increase in mask classification accuracy to 92.59%. These results highlight the potential of integrating MTL and custom loss functions in facial recognition. As face masks continue to be globally essential, DFN's integrated approach offers a fresh perspective in facial recognition studies. Furthermore, DFN paves the way for adaptive facial recognition systems, emphasizing the adaptability needed in our current era.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Somporn Ruang-on

Department of Creative Innovation in Science and Technology, Faculty of Science and Technology
Nakhon Si Thammarat Rajabhat University

1 Tambon Tha Ngio, Mueang Nakhon Si Thammarat, Nakhon Si Thammarat 80280, Thailand

Email: somporn_rua@nstru.ac.th

1. INTRODUCTION

Facial recognition, an esteemed pillar of computer vision, consistently positions itself at the vanguard of technological progression. The assimilation of advanced deep learning methodologies over recent years has facilitated its metamorphosis from basic image-matching paradigms [1] to complex feature extraction models [2], thereby rendering traditional manual engineering methodologies increasingly peripheral [3]. Recent studies [4], [5] have further highlighted the practical applications of face recognition in the context of smart city security and human emotion recognition, respectively. While these advanced systems exhibit remarkable proficiency in controlled settings, transitioning to real-world scenarios unveils a myriad of challenges. Factors such as inconsistent lighting, diverse ethnic backgrounds, age-related variations, and notable occlusions, especially face masks due to prevailing health concerns, accentuate the inherent imperfections of prevailing facial recognition frameworks [6].

In the expansive domain of facial recognition, face landmark detection crystallizes as a crucial preprocessing step, serving as a linchpin for a variety of applications [7], [8]. This foundational sub-discipline catalyzes the dynamism in emerging realms such as real-time facial expression recognition [9], immersive augmented reality ecosystems [10], and extends its significance to the security-centric domain of foolproof authentication mechanisms [11]. Recent trailblazing efforts encompass the work of Zhu *et al.* [12]

who proposed occlusion-adaptive deep networks to fortify facial landmark detection, Chandran *et al.* [13] who introduced attention-driven cropping for high-resolution facial landmark detection, and Li *et al.* [14] who pushed the boundaries with cascaded transformers for enhanced accuracy. These contributions not only signify the rapid advancements but also accentuate the evolving nature of this sub-discipline, showcasing a promising trajectory as it intersects with the broader domain of facial recognition, hinting at more sophisticated applications in the foreseeable future.

The ubiquitous use of face masks during the recent pandemic highlighted a significant gap: the absence of datasets tailored for landmark detection on masked faces. Such a deficiency undermines the performance of current models, emphasizing the urgency for methodologies that can adapt to these new challenges. A successful approach would merge the intricacies of facial landmark detection with face mask identification, leveraging the subtle nuances of facial contours and strategic landmark placement, even when partially obscured. While Gupta *et al.* [15] have made strides in mask detection, Ullah *et al.* [16] introduced the innovative DeepMaskNet model, bridging the gap between face mask detection and masked facial recognition. Doe and Smith developed two deep learning models, leveraging MobileNetv2 and a novel deep convolutional neural network (DCNN), to efficiently categorize mask usage into correctly worn, incorrectly worn, and not worn, using a Kaggle dataset for validation [17]. Additionally, Hdioud and Tirari [18] showcased the potential of deep learning for facial expression recognition of masked faces. Other notable works in the domain of mask detection include those by [19]–[21]. Altogether, these advances underscore the need for continuous evolution in face landmark detection techniques, which are pivotal in addressing the challenges presented by widespread mask usage and ensuring robust facial recognition in masked scenarios.

With face masks now entrenched in global societal norms, the fusion of these intertwined domains is essential for the subsequent phase of facial recognition advancements. Guided by these intricate challenges and the innovation potential, our research adopts a rigorous technical approach. We propose the use of semi-supervised learning techniques by jointly training a DCNN on both face landmark detection and face mask classification datasets. Drawing on the idea that knowledge from one domain can provide auxiliary information to another, our methodology leverages the shared feature space between face landmarks and mask classification. Preliminary observations suggest that this joint training not only enhances the granularity with which landmarks are detected on masked faces but also refines the accuracy and robustness of mask classifications. By coupling these tasks, we are essentially allowing our model to harness the mutual information between them, promoting a more generalized and effective learning process. Our initiative seeks to bridge the current gaps in the field by pioneering a method that optimally utilizes available data for enhanced performance on both tasks in real-world scenarios.

2. METHOD

In order to refine dual facial recognition, we amalgamate semi-supervised learning, utilizing both labeled and unlabelled datasets, to tackle the challenges posed by data paucity. This amalgamation dovetails with multi-task learning (MTL), where our innovative DualFaceNet (DFN) concurrently processes a gamut of facial attributes. We posit that shared feature spaces across these tasks markedly bolster task-specific performance. To further fortify our model, we infuse augmentation consistency loss, a mechanism that underpins model resilience to input fluctuations by mandating consistent outputs across diverse data augmentations. This synthesis establishes a rigorous foundation for our advanced facial recognition system, details of which will follow.

2.1. Model architecture: DualFaceNet

Our DFN was meticulously designed to adeptly manage dual objectives within facial recognition: discerning facial landmarks and classifying the presence of face masks. The neural network starts its processing pipeline by accepting an input of a $64 \times 64 \times 3$ color image. As the image progresses through the network, it traverses five 3×3 convolutional layers. Each of these convolutional layers employs a rectified linear unit (ReLU) as an activation function, chosen for its prowess in introducing non-linearity while also addressing the vanishing gradient problem. Following each convolution, there is a max-pooling layer condensing the spatial dimensions by half and thus, enhancing the model's translational invariance. The structure of these five primary layers is depicted as blue blocks in Figure 1. Notably, each block also defines the kernel size, delineated as $\text{width} \times \text{height} \times \text{input} \times \text{output}$, encapsulating the width, height, input channels, and output channels, respectively, of the kernels in these layers.

From this foundational structure, the neural network's architecture diverges into two specialized output pathways. The first pathway is dedicated to facial landmark detection. Transforming the captured spatial features, it employs a series of fully connected layers to output a vector of $2L$ real numbers. These numbers correspond to the 2D coordinates of L landmarks on the facial structure, with our standard configuration tailored for a granular 68 landmarks mapping.

Simultaneously, the second output pathway delves into the task of face mask classification. Distilling the learned features through its own set of fully connected layers, it culminates in producing a probability score. Leveraging the sigmoid activation function, this score offers a concise verdict on mask presence: scores veering toward 1 signify a mask worn correctly, while those approaching 0 denote otherwise. By amalgamating these dual outputs in a single architecture, as visually illustrated in Figure 1, our DFN crystallizes the essence of MTL, harmonizing two intertwined facial recognition tasks with seamless precision.

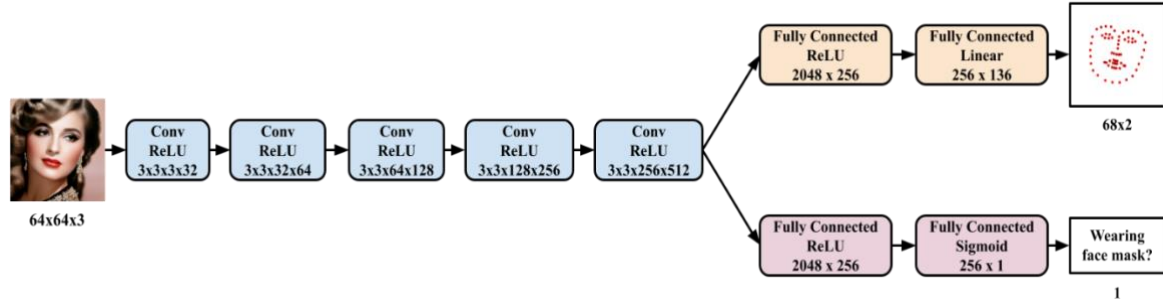


Figure 1. The architecture of our DFN

2.2. Loss functions

The essence of our training strategy is rooted in optimizing multiple loss functions concurrently, each tailored to a specific facet of our MTL paradigm. The overarching objective is to ensure robust and precise performance across both facial landmark detection and face mask classification. We delineate the various loss functions and their roles in the training process as follows.

2.2.1. Face landmark loss

This foundational landmark loss function emphasizes the accuracy of facial landmark localization. It computes the discrepancy between the predicted landmarks and the ground truth using mean absolute error (MAE), aiming to linearly minimize this differential. Specifically, the landmark loss is defined in (1).

$$L_{Landmark} = \frac{1}{NL} \sum_i^N \sum_j^L |l_{ij} - \hat{l}_{ij}| \quad (1)$$

where N is the number of data points, L is the number of total landmarks in each facial image, l_{ij} are ground-truth landmarks locations, and \hat{l}_{ij} are the predicted landmarks by the model.

2.2.2. Face mask loss

Central to the undertaking of face mask classification, this loss metric evaluates the discrepancy between the predicted mask-wearing status, signifying whether a mask is worn correctly or not, and its actual status as delineated in the ground truth utilizing binary cross entropy. Refer to (2) for a more detailed definition of this metric. Where N is the number of data points, y_i are ground-truth wearing mask labels, and \hat{y}_i are the wearing mask predictions from the model.

$$L_{Mask} = \frac{1}{N} \sum_i^N y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \quad (2)$$

2.2.3. Face landmark consistency loss

This loss metric is designed to foster consistent facial landmark predictions across varying renditions of the same image, such as original and augmented versions. By minimizing discrepancies in landmark predictions, the model's stability and reliability are significantly enhanced. This consistency can be quantitatively defined through the MAE as expressed in (3).

$$L_{LandmarkConsistent} = \frac{1}{NL} \sum_i^N \sum_j^L |\hat{l}_{ij}^{orig} - inv(\hat{l}_{ij}^{aug})| \quad (3)$$

Where N is the number of data points, L is the number of total landmarks in each facial image, \hat{l}_{ij}^{orig} are the predicted landmarks of the original images by the model, \hat{l}_{ij}^{aug} are the predicted landmarks of the augmented images, and $inv()$ is an inverse image transformation of the applied augmentations.

2.2.4. Face mask consistency loss

Serving a parallel purpose to the landmark consistency loss, yet tailored for medical face mask classification, this loss function strives to ensure that mask predictions exhibit consistency across diverse representations of the same image, as quantified in (4).

$$L_{MaskConsistent} = \frac{1}{N} \sum_i |\hat{y}_i^{orig} - \hat{y}_i^{aug}| \quad (4)$$

where N is the number of data points, \hat{y}_{ij}^{orig} are mask prediction of the original images, and \hat{y}_{ij}^{aug} are the mask predictions of the augmented images from the model.

2.2.5. Total loss

The total loss of the training procedure is derived from a linear combination of the aforementioned loss metrics. This amalgamation serves as a pivotal measure, directing the optimization of DFN towards enhanced performance in both facial landmark detection and mask classification tasks. Specifically, the integration of these individual loss components into a single total loss metric is articulated in (5), aiming to concurrently minimize the discrepancies in facial landmark predictions and mask classification across diverse image representations.

$$L_{Total} = L_{Landmark} + L_{Mask} + L_{LandmarkConsistent} + L_{MaskConsistent} \quad (5)$$

2.3. Model training strategy

Our training paradigm for the DFN architecture is illustrated in Figure 2, showcasing our dual-dataset strategy designed for multitasking across facial landmark detection and face mask classification tasks, each sourcing data from a distinct dataset. The facial landmark detection dataset encompasses a diverse collection of facial images. The primary objective of this dataset is to train the model to detect and accurately map facial landmarks. The corresponding loss metric, denoted as $L_{Landmark}$ in (1), is structured to ensure precise identification and localization of these landmarks across a variety of facial structures. Conversely, the face mask classification dataset serves as an auxiliary yet crucial dataset, containing images that distinctly demonstrate individuals either wearing face masks correctly or not. The nuanced task posed by this dataset is to train the model to discern the presence or absence of face masks. The affiliated loss function, denoted as L_{Mask} as defined in (2), focuses on maximizing the accuracy of mask classification, thus working in tandem with the landmark detection task to ensure a holistic, robust performance of our DFN model across these intertwined facial recognition tasks.

To further enhance model generalization, data augmentation techniques are uniformly applied to both datasets. These manipulations inject realistic variability into the data, thereby promoting robust learning. Critically, the augmentation consistency losses $L_{LandmarkConsistent}$ and $L_{MaskConsistent}$ in (3) and (4) ensure the model's predictions remain stable despite these transformations, cementing its resilience. During the training cycle, the total loss in (5) is continuously evaluated and optimized. This orchestrated synergy between the two datasets, augmented by their individual loss functions, ensures that DFN is finely tuned to excel in both facial landmark detection and face mask classification.

2.4. Data augmentation techniques

In the domain of deep learning, data augmentation is crucial for bolstering the generalization capabilities of models, particularly when there's a scarcity of training data. Throughout the experiments, we implemented a set of augmentation techniques tailored to address the distinctive challenges of facial recognition and landmark detection. These techniques ensure the model's robustness against a wide array of real-world scenarios. Firstly, we utilized random cropping and re-scaling to emulate variations in face sizes and positions, ensuring the model's adaptability to various face placements and scales. Secondly, random rotation was employed to account for potential tilts in faces, rotating images within a range of ± 30 degrees. Thirdly, recognizing that lighting can vary drastically across environments, we randomly adjusted image brightness and contrast to ensure model resilience against such fluctuations. Fourthly, horizontal flipping was incorporated, flipping images at a 50% probability rate, which not only expands the effective dataset size but also confirms the model's invariance to face orientation. It's essential to adjust facial landmark annotations correspondingly for any flipped images. By amalgamating these augmentation strategies, we have enhanced DFN's training on a

diverse array of facial scenarios, bolstering its generalization capabilities. This rigorous augmentation approach was pivotal in achieving the impressive performance metrics recorded in our evaluations.

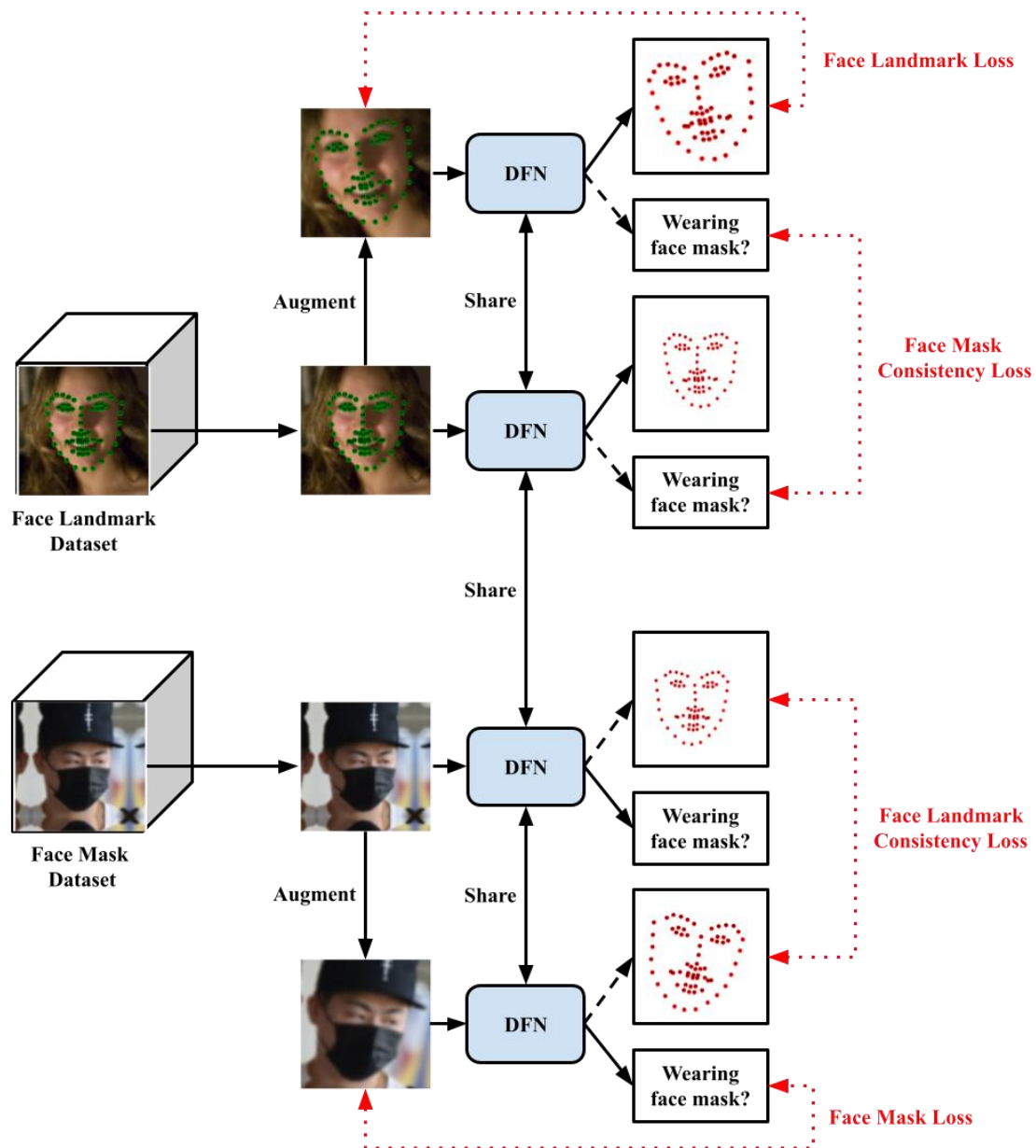


Figure 2. Overview training strategy of our DFN

2.5. Implementation details

Implementing our multi-task facial recognition involves training the DFN to handle the dual tasks of facial landmark detection and face mask classification. All experiments were conducted using Python 3 and the Keras framework for its versatility and efficiency in handling deep learning tasks. The input facial images were normalized to fall between 0 and 1, achieved by dividing each pixel value by 255. Training complex deep learning models, especially for tasks like multi-task facial recognition, can be both time-consuming and resource-intensive. One of the challenges faced during this process is the tuning of hyper-parameters. To streamline our training process, we opted for practical default values for certain hyper-parameters. We utilized a batch size of 64 and employed the Adam optimizer [22], which has shown consistent performance in training intricate neural networks. The models were trained for 200 epochs, where a learning rate of 1×10^{-3} was used during the first 100 epochs, and 1×10^{-5} was used for the rest. Furthermore, we set the weight decay at 5×10^{-4} .

3. RESULT AND DISCUSSION

In this section, we will dive into the experiments we conducted to examine the DFN, especially in the areas of facial landmark detection and mask classification. We will start by discussing the datasets we used, shedding light on what they contain and why they were chosen. Next, we will touch on the metrics we used to measure how well our model performs. We will also compare our model with other models. Finally, we will wrap up with an analysis of our results, giving you a clear picture of what we found and what it means for facial recognition research.

3.1. Datasets

Our approach to facial landmark detection and face mask classification sought to capitalize on the precision of available annotations while eliminating the need for extensive joint labeling. By harnessing datasets labeled independently for each task, we could focus on the nuances and specificities inherent to each domain. This approach not only streamlined our model training and evaluation processes but also highlighted the potential of MTL when tasks can be addressed without the complexities and overheads of concurrent annotations. This strategic utilization of pre-existing, task-specific datasets underscores the efficiency and adaptability of our methodology.

3.1.1. The 300W dataset

The 300W dataset stands as a pivotal contribution in the domain of facial landmark detection, setting a benchmark that the scientific community fervently adheres to. It offers a diverse collection of facial images, methodically curated to rigorously test facial landmark detection algorithms. The dataset is orchestrated into multiple subsets, each meticulously tailored to reflect specific challenges encompassed within real-world scenarios, such as diverse lighting conditions, a broad spectrum of facial expressions, and varying degrees of occlusions. Owing to its meticulous curation and versatility, the 300W dataset offers invaluable insights for training and evaluation processes related to facial landmark detection [23]. The dataset encompasses 3,148 training images alongside 600 test images, thus offering a substantial volume of data for rigorous analysis. To offer a visual insight into the rich diversity and granularity of this dataset, refer to Figure 3(a), which illustrates sample facial images from the 300W dataset along with their corresponding landmarks.

3.1.2. Face mask classification dataset

For the face mask classification endeavor, our choice was a detailed dataset as presented by Su *et al.* [24]. This dataset owes its inception to the esteemed works of Wang *et al.* [25] and the MAFA datasets, credited to Ge *et al.* [26]. Each image in this collection is harmoniously resized to a consistent resolution of 224×224 pixels, ensuring uniform input for subsequent analyses. The dataset delineates masks into two distinct categories. The first, qualified masks (OK-mask), encompasses 1,361 images, predominantly highlighting N95 masks and disposable medical variants, which are globally recognized for their superior filtration capabilities. Conversely, the unqualified masks (NG-mask) segment contains 1,880 images, portraying masks that fall short of medical protection standards, such as sponge masks, cloth variants, scarves, and other unconventional facial coverings. In total, the dataset boasts a formidable compilation of 3,241 images, a select few of which are depicted in Figure 3(b). The dataset was divided into a training set of 2,593 images and a test set comprising 648 images. Leveraging this open-access dataset allowed us to execute rigorous mask classification experiments and draw substantial inferences.

3.2. Metrics

In the domain of facial recognition and landmark detection, a rigorous and precise evaluation of model performance is indispensable. This evaluative process, underpinned by quantifiable metrics, not only substantiates the integrity of the research but also elucidates potential avenues for enhancement. Among the plethora of evaluation metrics, two have emerged as particularly salient in this context: accuracy and the interocular normalized mean error (INME).

3.2.1. Accuracy

Accuracy is a cornerstone metric in machine learning and classification endeavors. It quantifies the proportion of instances correctly identified by a model in relation to the entire dataset. Its simplicity and directness render it a fundamental tool in the assessment repertoire. However, it is imperative to approach this metric with circumspection, particularly when dealing with datasets that exhibit class imbalances. The mathematical representation of the accuracy is illustrated in (6).

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Number of total predictions}} \quad (6)$$



Figure 3. Sample facial images and their annotation from the (a) 300W dataset and (b) face mask classification dataset

3.2.2. Interocular normalized mean error

INME offers a more intricate assessment tailored specifically for facial landmark detection. This metric measures the average discrepancy between predicted and actual landmark positions, subsequently normalizing this value by the interocular distance defined as the distance between the two most exterior points of the eyes. This normalization process ensures a scale-invariant evaluation. The formulation for INME is defined in (7).

$$INME = \frac{1}{N} \sum_i^N \frac{\sqrt{\sum_j^L (l_{ij} - \hat{l}_{ij})^2}}{D_i} \quad (7)$$

Where N is the number of data points, L is the number of total landmarks in each facial image, l_{ij} are ground-truth landmarks locations, \hat{l}_{ij} are the predicted landmarks by the model, and D_i is the distance between the outer eye corners of each image.

3.3. Methods

In this subsection, we delve into the methodologies employed in our experiments, specifically: the landmark baseline, face mask baseline, MTL, and DFN. For each of these methods, we provide an in-depth analysis of their training and validation performances. This comprehensive examination offers a clear perspective on the efficacy and nuances of each approach in the context of our study.

3.3.1. Landmark baseline

Landmark baseline refers to a method with the same architecture as DFN but without the face mask output branch. As this method was trained solely on the landmark dataset, it can only be evaluated with the INME. It serves as a landmark detection baseline compared to our jointly trained method. The training and validation INMEs during the training of 200 epochs of the landmark baseline are shown in Figure 4(a). The figure shows a discernible disparity between the training INME and the validation INME. In the initial 100 epochs, both metrics decline at an exponential rate, appearing to stabilize after the 70th epoch. After reducing the learning rate to 1×10^{-5} up to the 100th epoch, there was a marked decrease in the INME values. They then stabilize, with the training INME settling at 2.13 and the validation INME at 5.61.

3.3.2. Face mask baseline

Similarly, the face mask baseline refers to a method with the same architecture as DFN, and it was also trained on the face mask dataset only. As a result, its performance can only be measured with face mask

accuracy. This approach acts as a face mask classification baseline for our proposed method. The training and validation face mask classification accuracy during training of 200 epochs of the face mask baseline were shown in Figure 4(b). From the figure, a clear disparity is evident between the training accuracy and the validation accuracy for face mask classification. Throughout the initial epochs, both accuracies demonstrate a sharp upward trend. However, post a certain point, while the training accuracy continues its ascent, reaching an impressive 99.99%, the validation accuracy appears to plateau, settling at 89.35%. This divergence underscores the challenges of generalization and the nuances of the validation set compared to the training data.

3.3.3. Multi-task learning

MTL unfolds as a pragmatic strategy wherein a singular model is trained concurrently on multiple interrelated tasks, thereby harnessing shared information to foster enhanced generalization. In our investigative endeavor, we employed MTL to harmoniously navigate through dual objectives: the detection of facial landmarks and the classification of face masks. This stratagem capitalizes on the identical architectural foundation as our DFN. However, it's pertinent to underscore that despite sharing its core strength with our DFN, MTL, in the absence of consistency loss, does not impose any artificial similarity or coherence between augmented images, which could be a pivotal aspect in certain scenarios. The empirical journey of training and validation within the realms of facial landmark detection and face mask classification has been pictorially represented in Figures 4(c) and 4(d), respectively. These figures encapsulate the evolutionary trajectory of model accuracy across these intertwined tasks, offering a visual narrative of the model's performance.

In Figure 4(c), which showcases the results for face landmark detection from MTL, we observe a trajectory akin to our baseline model but with nuanced differences. The training INME demonstrates a swift descent, eventually stabilizing at a commendable 2.37, indicative of the model's adeptness in landmark detection. The validation INME follows a similar pattern, though it plateaus slightly higher at 5.90. Transitioning to Figure 4(d), which focuses on face mask classification from MTL, the results echo the trends seen in Figure 4(c). The training accuracy exhibits a robust climb, peaking at an impressive 100.00%. However, the validation accuracy, while initially tracking the training accuracy closely, begins to diverge as epochs progress, culminating at 91.98%.

3.3.4. DualFaceNet

Our DFN is an innovative technique that harmoniously fuses multiple sources of information for enhanced performance by integrating insights from both facial landmark detection and face mask classification. While the architectural foundation of DFN parallels that of MTL, DFN introduces consistency losses to ensure robustness against variations, especially in augmented scenarios. The performance metrics for face landmark detection and face mask classification using DFN are illustrated in Figures 4(e) and 4(f), respectively.

In Figure 4(e), which presents the results for face landmark detection, there was a marked distinction compared to prior models. The training INME starts with a swift decline, indicative of DFN's rapid learning capability, and eventually plateaus at 2.49, underlining the network's precision in detecting facial landmarks. The validation INME, while charting a similar course, stabilizes at a slightly elevated 5.42. Switching our attention to Figure 4(f), which illustrates the face mask classification results, the patterns are reminiscent of those in Figure 4(e) but with their unique characteristics. The training accuracy accelerates sharply, reaching a near-perfect 100%. In contrast, the validation accuracy, although beginning on a promising note, finds its equilibrium at 92.59%.

3.4. Methods comparison

The efficacy of facial recognition models is intrinsically linked to their performance metrics. In this section, we compare the performances in terms of INME and face mask accuracies among the face landmark baseline, face mask baseline, MTL, and DFN. The Table 1 summarizes the performance of different approaches.

The results table illuminates key insights into the model's performances. Baseline models, tailored for either landmark detection or face mask classification, provide foundational benchmarks with the face landmark model reporting an INME of 5.61 and the face mask model attaining an accuracy of 89.35%. The transition to MTL, encompassing simultaneous training for both tasks, leads to a minor uptick in landmark error to 5.90, yet face mask classification accuracy sees a commendable leap to 91.98%. For DFN, integrating consistency loss further sharpens these metrics, bringing down landmark error to 5.42 and boosting mask accuracy to 92.59%. Our result is comparable to the state of the art presented in [17]. This progression emphasizes the transformative potential of MTL in facial recognition, particularly when enhanced with additional specialized loss functions such as consistency loss. To elucidate the training dynamics of the models, Figures 5(a) and 5(b) respectively present a comparative view of the validation INME and face mask accuracy across the training epochs.

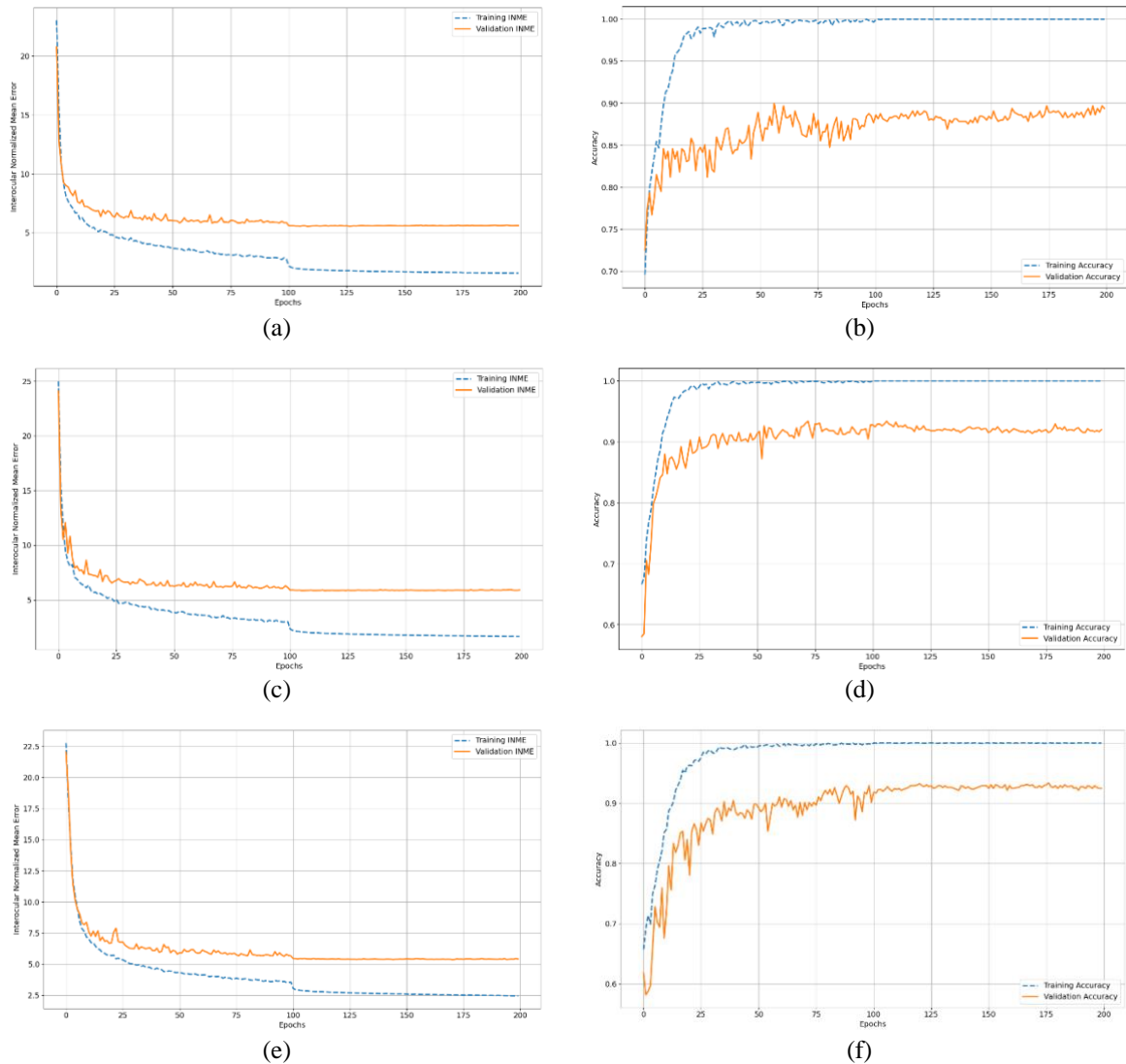


Figure 4. Training and validation of face landmark INME and face mask accuracy of different models: (a) landmark INME from baseline, (b) face mask accuracy from baseline, (c) landmark INME from MTL, (d) face mask accuracy from MTL, (e) landmark INME from DFN (Our), and (f) face mask accuracy from DFN (our)

Table 1. Method comparison for facial landmark detection and face mask classification

Methods	INME↓	Accuracy (%)
Landmark baseline	5.61	-
Face mask baseline	-	89.35
MTL	5.90	91.98
DFN (our)	5.42	92.59

Figure 5(a) provides a visual narrative of the validation INME trends across the training epochs for the different models, painting a vivid picture that complements the tabulated results. The curve for the face landmark baseline serves as the foundational benchmark, tracing a path indicative of its inherent strengths in facial landmark detection. This behavior, in harmony with its reported INME of 5.61, establishes the performance standard against which the other models are evaluated. Transitioning to the MTL curve, we observe an intriguing pattern. Although one might expect gains from simultaneous training on multiple tasks, the curve reveals a slightly higher plateau at an INME value, corresponding to its tabulated 5.90. This visual representation underscores the notion that MTL, in this context, might not always lead to enhanced performance, even faltering slightly compared to the specialized baseline. Lastly, the trajectory of the DFN emerges as a beacon of promise. With its rapid descent and subsequent stabilization, the curve visually

echoes its superior tabulated INME of 5.42. This affirms DFN's proficiency in facial landmark detection, particularly when enhanced with augmented consistency loss.

Figure 5(b) maps the validation accuracy for face mask classification across distinct models and training epochs. Starting with the face mask baseline, its trajectory serves as a foundational reference. The steady climb it portrays resonates with its tabulated accuracy of 89.35%, establishing a baseline metric that more complex models aim to surpass. Progressing to the MTL curve, we witness a heartening surge. Contrasting the baseline, MTL's curve showcases a more robust ascent, settling at a plateau that mirrors its reported accuracy of 91.98%. This ascent underscores the advantages of simultaneous training on intertwined tasks, as MTL successfully bridges the gap between specialized singular models and more intricate multi-task frameworks. However, the zenith of performance is captured by the DFN trajectory. Beginning in tandem with MTL, a pivotal moment transpires just after epoch 100 where DFN's trajectory begins its overtaking maneuver. This surge, culminating in a pinnacle reflective of its superior tabulated accuracy of 92.59%, confirms DFN's supremacy in face mask classification. The integration of consistency loss offers DFN this edge, allowing it not only to surpass the baseline but also to outpace MTL.

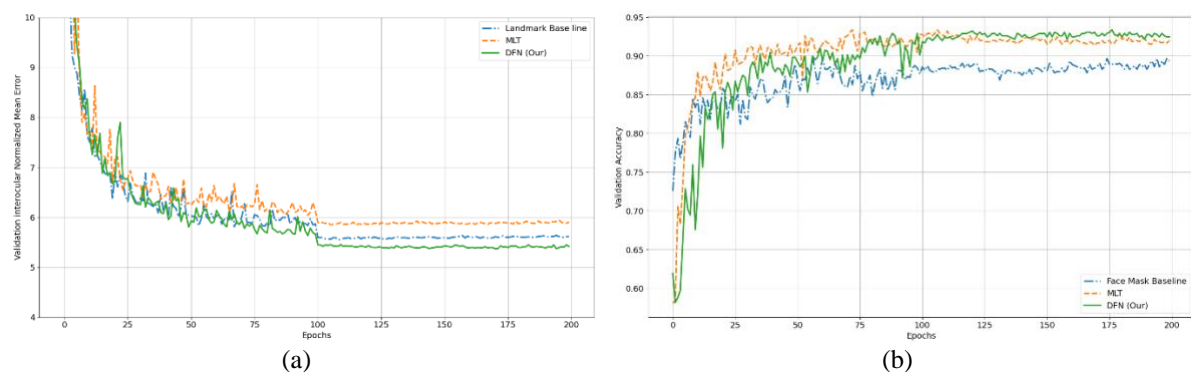


Figure 5. The comparison of validation face landmark INME and face mask accuracy of different methods: (a) validation face landmark INME, and (b) validation face mask accuracy

4. CONCLUSION

The dynamic realm of facial recognition is at a pivotal juncture, with real-world challenges necessitating adaptive methodologies. Our research ventured into this domain, introducing DFN, a groundbreaking approach synergistically merging facial landmark detection and face mask classification. By capitalizing on MTL and consistency loss, DFN transcends traditional single-task models in performance. Comprehensive evaluations, encompassing diverse datasets and intricate metrics, attest to DFN's prowess, particularly in navigating occlusions such as masks. As face masks solidify their presence in global society, DFN's fusion of landmark detection and mask classification becomes increasingly vital for future facial recognition advancements. Anticipating the future, we envision integrating real-time video analysis with DFN to enhance surveillance and security mechanisms. Further enrichments could arise from adding tasks to DFN, such as emotion detection or age estimation. Also, testing DFN on larger and more varied datasets will be pivotal to gauging its scalability and robustness. By relentlessly pushing these frontiers, we aim to sculpt new benchmarks in the ever-evolving world of facial recognition.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the use of service and facilities of the Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University. This study receives funding from the Coordinating Center for Thai Government Science and Technology Scholarship Students (CSTS) and National Science and Technology Development Agency (NSTDA), under the Research Grant Scheme JRA-CO-2565-17792-TH.

REFERENCES




- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
- [3] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.
- [4] G. B. Praveen and J. Dakala, "Face recognition: challenges and issues in smart city/environments," *2020 International*

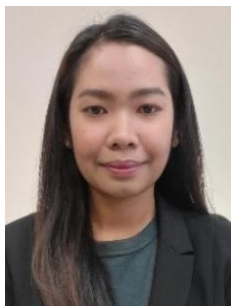
DualFaceNet: augmentation consistency for optimal facial landmark detection and ... (Kritaphat Songsri-in)




- Conference on COMMunication Systems and NETWORKS, COMSNETS 2020*. IEEE, pp. 791–793, 2020, doi: 10.1109/COMSNETS48256.2020.9027290.
- [5] R. Amimi, A. Radgui, and H. I. E. H. El, “A Survey of smart classroom: concept, technologies and facial emotions recognition application,” *Lecture Notes in Networks and Systems*, vol. 544. Springer International Publishing, pp. 326–338, 2023, doi: 10.1007/978-3-031-16075-2_23.
 - [6] C. Libby and J. Ehrenfeld, “Facial recognition technology in 2021: masks, bias, and the future of healthcare,” *Journal of Medical Systems*, vol. 45, no. 4, Feb. 2021, doi: 10.1007/s10916-021-01723-w.
 - [7] K. S. -In, G. Trigeorgis, and S. Zafeiriou, “Deep and deformable: convolutional mixtures of deformable part-based models,” *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*. IEEE, pp. 218–225, 2018, doi: 10.1109/FG.2018.00040.
 - [8] Y. Wu and Q. Ji, “Facial landmark detection: a literature survey,” *International Journal of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019, doi: 10.1007/s11263-018-1097-z.
 - [9] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, “Facial expression analysis under partial occlusion: a survey,” *ACM Computing Surveys*, vol. 51, no. 2, pp. 1–49, 2018, doi: 10.1145/3158369.
 - [10] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.
 - [11] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: a unified embedding for face recognition and clustering,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015, doi: 10.1109/CVPR.2015.7298682.
 - [12] M. Zhu, D. Shi, M. Zheng, and M. Sadiq, “Robust facial landmark detection via occlusion-adaptive deep networks,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 3481–3491, 2019, doi: 10.1109/CVPR.2019.00360.
 - [13] P. Chandran, D. Bradley, M. Gross, and T. Beeler, “Attention-driven cropping for very high resolution facial landmark detection,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 5860–5869, 2020, doi: 10.1109/CVPR42600.2020.00590.
 - [14] H. Li, Z. Guo, S. M. Rhee, S. Han, and J. J. Han, “Towards accurate facial landmark detection via cascaded transformers,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 4166–4175, 2022, doi: 10.1109/CVPR52688.2022.00414.
 - [15] P. Gupta, V. Sharma, and S. Varma, “A novel algorithm for mask detection and recognizing actions of human,” *Expert Systems with Applications*, vol. 198, Jul. 2022, doi: 10.1016/j.eswa.2022.116823.
 - [16] N. Ullah, A. Javed, M. A. Ghazanfar, A. Alsufyani, and S. Bourouis, “A novel DeepMaskNet model for face mask detection and masked facial recognition,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 9905–9914, Nov. 2022, doi: 10.1016/j.jksuci.2021.12.017.
 - [17] A. A. Abdulmunem, N. D. A. -Shakarchy, and M. S. Safoq, “Deep learning based masked face recognition in the era of the COVID-19 pandemic,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 1550–1559, 2023, doi: 10.11591/ijece.v13i2.pp1550-1559.
 - [18] B. Hdoud and M. E. H. Tirari, “Facial expression recognition of masked faces using deep learning,” *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 921–930, 2023, doi: 10.11591/ijai.v12.i2.pp921-930.
 - [19] C. X. Ge, M. A. As’ari, and N. A. J. Suffri, “Multiple face mask wearer detection based on YOLOv3 approach,” *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 384–393, 2023, doi: 10.11591/ijai.v12.i1.pp384-393.
 - [20] B. U. H. Sheikh and A. Zafar, “RRFMDS: rapid real-time face mask detection system for effective COVID-19 monitoring,” *SN Computer Science*, vol. 4, no. 3, pp. 288, 2023, doi: 10.1007/s42979-023-01738-9.
 - [21] S. Susanto, F. A. Putra, R. Analia, and I. K. L. N. Suciningtyas, “The face mask detection for preventing the spread of COVID-19 at politeknik negeri batam,” *Proceedings of ICAE 2020 - 3rd International Conference on Applied Engineering*. IEEE, 2020, doi: 10.1109/ICAES0557.2020.9350556.
 - [22] D. P. Kingma and J. L. Ba, “Adam: a method for stochastic optimization,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
 - [23] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, “300 faces in-the-wild challenge: the first facial landmark localization challenge,” *The IEEE International Conference on Computer Vision*. IEEE, pp. 397–403, 2013, doi: 10.1109/ICCVW.2013.59.
 - [24] X. Su, M. Gao, J. Ren, Y. Li, M. Dong, and X. Liu, “Face mask detection and classification via deep transfer learning,” *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 4475–4494, 2022, doi: 10.1007/s11042-021-11772-5.
 - [25] Z. Wang, B. Huang, G. Wang, P. Yi, and K. Jiang, “Masked face recognition dataset and application,” *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 2, pp. 298–304, 2023, doi: 10.1109/TBIOM.2023.3242085.
 - [26] S. Ge, J. Li, Q. Ye, and Z. Luo, “Detecting masked faces in the wild with LLE-CNNs,” *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, IEEE, pp. 426–434, 2017, doi: 10.1109/CVPR.2017.53.

BIOGRAPHIES OF AUTHORS






Kritaphat Songsri-in    finished M.Eng. and Ph.D. in computing from Imperial College London in 2011 and 2020, respectively. Currently, he is an Assistant Professor in the Department of Computer Science at Nakhon Si Thammarat Rajabhat University, Thailand. His research interests include machine learning, deep learning, and computer vision. He has published in and is a reviewer for multiple international conferences and journals such as IEEE Transactions on Image Processing and IEEE Transactions on Information Forensics & Security. He was a recipient of the Royal Thai Government Scholarship covering his undergraduate and postgraduate degrees in 2010. He received the Best Student Paper Awards at the IEEE 13th International Conference for Automatic Face and Gesture Recognition (FG2018) and the 6th National Science and Technology Conference (NSCIC2021). In 2021, his Ph.D. thesis received an award from the National Research Council of Thailand (NRCT). He can be contacted at email: kritaphat_son@nstru.ac.th.






Munlika Rattaphun    received the B.S. degree in computer science from Thaksin University, Songkhla, Thailand, in 2009, the M.S. degree in computer science from Prince of Songkla University, Songkhla, Thailand, in 2011, and the Ph.D. degree in computer science and information engineering from National Chiayi University, Chiayi, Taiwan, in 2022. She is currently a lecturer at the Department of Computer Science, Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Nakhon Si Thammarat, Thailand. Her current research interests include machine learning, nearest-neighbor search, and recommender systems. She can be contacted at email: munlika_rat@nstru.ac.th.



Sopee Kaewchada    received the B.Sc. degree in computer science from Rajabhat Phetchaburi Institute, Thailand, in 1997 the M.S. degree in management of information technology from Walailak University, Thailand, in 2003, and the Ph.D. degree in Creative Innovation in Science and Technology, Nakhon Si Thammarat Rajabhat University, Thailand, in 2023. Currently, she is an Assistant Professor at the Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Thailand. She can be contacted at email: sopee_kae@nstru.ac.th.



Somporn Ruang-On    received the B.Sc. degree in computer science from Rajabhat Phetchaburi Institute, Thailand, in 1995, the M.Sc. degree in information technology from Sripatum University, Thailand, in 2003, and Ph.D. degree in Quality information technology from Phetchaburi Rajabhat University, in 2013. Currently, he is an Assistant Professor at the Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Thailand. He can be contacted at email: somporn_rua@nstru.ac.th.